Semantic Concept Discovery Over Event Databases

May 24, 2017 01:11 Draft (Confidential – Please Do Not Distribute)

Oktie Hassanzadeh
IBM Research
hassanzadeh@us.ibm.com

Shari Trewin
IBM Research
trewin@us.ibm.com

Alfio Gliozzo
IBM Research
gliozzo@us.ibm.com

ABSTRACT

In this paper, we study the problem of identifying entities of certain types (specifically, people and organizations) relevant to a given analysis question with the goal of assisting a human analyst in researching that question. We consider a case where we have a large event database describing events and their associated news articles along with meta-data describing various event attributes such as people and organizations involved and the topic of the event. We describe the use of semantic technologies in question understanding and deep analysis of the event database, and show a detailed evaluation of our proposed concept discovery techniques using reports from Human Rights Watch organization and other sources. Our study finds that combining our neural network based semantic term embeddings over structured data with an index-based method can significantly outperform either method alone.

KEYWORDS

Concept Discovery, Event Databases, Semantic Data Integration, Semantic Embeddings

ACM Reference format:
Oktie Hassanzadeh, Shari Trewin, and Alfio Gliozzo. 2017. Semantic Concept Discovery Over Event Databases. In *Proceedings of ACM International Conference on Information and Knowledge Management (CIKM), Singapore, November 2017 (CIKM'17),* 9 pages.
https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Analysts are often tasked with preparing a comprehensive, accurate, and unbiased report on a given topic. The first step in preparing such a report is a daunting discovery task that requires researching through a massive amount of information. Information sources can have large volume, variety, varying veracity, and velocity - the common characteristics of the so-called Big Data sources. Many times the analysis requires a deep understanding of various kinds of historical and ongoing *events* that are reported in the media. To enable better analysis of events, there exist several *event databases* containing structured representations of events extracted from news articles. Examples include GDELT [22], ICEWS [16], and EventRegistry [21]. These event databases have been successfully used to perform various kinds of analysis tasks, e.g., forecasting societal events [27]. However, there has been little work on the discovery aspect of the analysis, that results in a gap between the information requirements and the available data, and potentially a biased view of the available information.

In this paper, we present a framework for concept discovery over event databases using semantic technologies. Unlike existing concept discovery solutions that perform discovery over text documents and in isolation from the remaining data analysis tasks [23, 33], our goal is providing a unified solution that allows deep understanding of the same data that will be used to perform other analysis tasks (e.g., hypothesis generation [31, 32] or building models for forecasting [20, 27]). Figures 1 & 2 show different views of our system's UI that is built using our concept discovery framework APIs. The analyst can enter a natural language question or a set of concepts, and retrieve collections of relevant concepts identified and ranked using different discovery algorithms described in Section 3. A key aspect of our framework is the use of semantic technologies. In particular:

- A unified view over multiple event databases and a background RDF knowledge base is achieved through semantic link discovery and annotation.
- Natural language or keyword query understanding is performed through mapping of input terms to the concepts in the background knowledge base.
- Concept discovery and ranking is performed through neural network based semantic term embeddings.

In what follows, we first describe the overall framework and its various components. We then describe the algorithms used for concept discovery and ranking. In Section 4, we present the methodology and results of our evaluation using a ground truth built from a large corpus of reports written by human experts.

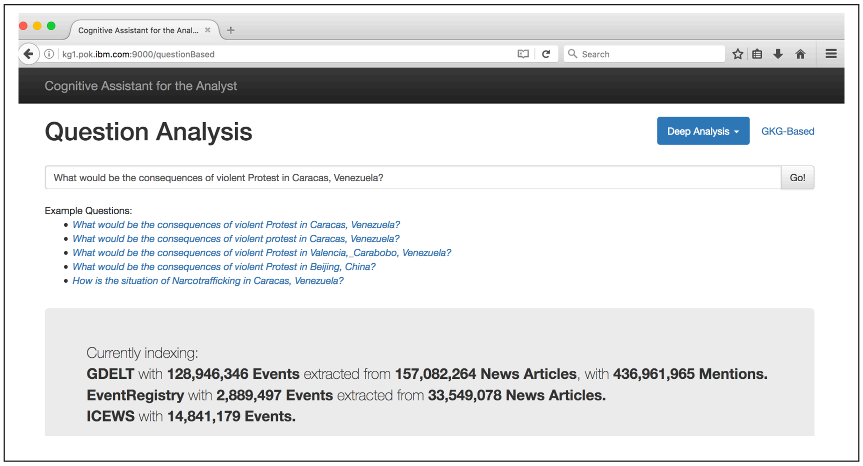
2 CONCEPT DISCOVERY FRAMEWORK

Figure 3 shows the architecture of our system. The system takes in as input a set of event databases and RDF knowledge bases and provides as output a set of APIs that provide a unified retrieval mechanism over input data and knowledge bases, and an interface to a number of concept discovery algorithms. In what follows, we describe the input sources and each of the components in detail.

2.1 Event Data & Knowledge Sources

Event databases are structured records describing various kinds of societal, political, or economic events. While event extraction from text is a well-studied topic in the NLP literature [17, 19] with a dedicated track at the annual Text Analysis Conference (TAC) [10], there are only a few publicly available large-scale event databases.

CIKM'17, November 2017, Singapore
© 2017 Copyright held by the owner/author(s).
ACM ISBN 978-x-xxxx-xxxx-x/YY/MM.
https://doi.org/10.1145/nnnnnnn.nnnnnnn


Cognitive Assistant for the Anal...
kg1.pok.ibm.com:9000/questionBased
Search
Cognitive Assistant for the Analyst
Deep Analysis
GKG-Based
Go!

Question Analysis

What would be the consequences of violent Protest in Caracas, Venezuela?

Example Questions:

- *What would be the consequences of violent Protest in Caracas, Venezuela?*
- *What would be the consequences of violent protest in Caracas, Venezuela?*
- *What would be the consequences of violent Protest in Valencia,_Carabobo, Venezuela?*
- *What would be the consequences of violent Protest in Beijing, China?*
- *How is the situation of Narcotrafficking in Caracas, Venezuela?*

Currently indexing:
GDELT with **128,946,346 Events** extracted from **157,082,264 News Articles**, with **436,961,965 Mentions.**
EventRegistry with **2,889,497 Events** extracted from **33,549,078 News Articles.**
ICEWS with **14,841,179 Events.**

Figure 1: Question Analysis UI - Main Page

The input of these event databases is a large corpus of news articles that are either gathered from various news sources (e.g., news agencies and other proprietary sources) or crawled from the Web (e.g., as done by Google News [7] or AlchemyData News [1]). The output is structured records (i.e., relational data tables) describing various features of the identified events.

2.1.1 GDELT. The Global Data on Events, Location, and Tone (GDELT) project [22] claims to be "the largest, most comprehensive, and highest resolution open database of human society ever created". GDELT data contains three databases. GDELT Event database is among a class of event databases that provide *coded* event data. The coding is based on a popular scheme using the CAMEO (Conflict And Mediation Event Observations) coding framework [29]. In this coding system, each event consists of a maximum of two *actors* and an *action*. Actors and actions are coded based on a hierarchy provided by CAMEO. In addition to coded actors and actions, GDELT includes other features of the events such as the date of the event (the date first reported), the source article (the first article mentioning the event), the number of articles and news sources mentioning the event, numerical scores reflecting the "tone" of the articles mentioning the event and other similar features, and geographical coordinates of the actors and the action. The second database provided by GDELT is the Global Knowledge Graph (GKG). Unlike Event database records that represent events, records in the GKG database describe the source articles of the events. Each record provides a comprehensive annotation of the articles with various numerical features such as tone and a set of measures referred to as Global Content Analysis Measures (GCAM) [8], in addition to annotations with several dictionaries of persons, organizations, and "themes". The third database GDELT provides is the Mentions database which connects event records with GKG article records. The most recent version of GDELT data is updated daily and includes historical data since February 2015, while the older version of the event database (GDELT 1.0) includes events since January 1979. At the time of this writing, we have ingested 128,946,346 Event records, 157,082,264 GKG records, and 436,961,965 Mention records.

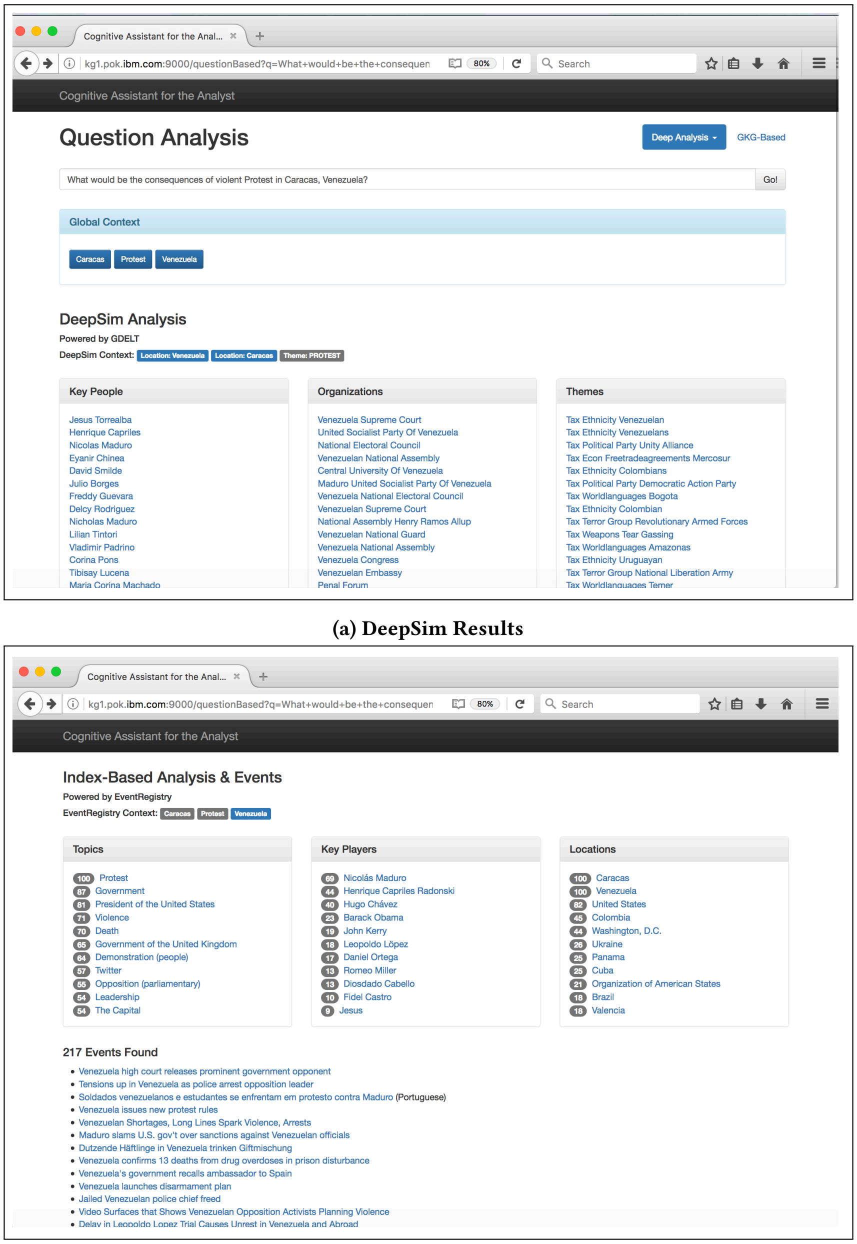
2.1.2 ICEWS. Integrated Conflict Early Warning System [12, 16, 35] provides a coded event database similar to the GDELT Events database. ICEWS event records describe features of source and target actors including their name, "sector", and country, features of the action including date (first reported), source, a short text description, and geographical descriptions. A recent version of the data also includes CAMEO codes for actions. We have ingested the most recent publicly available ICEWS data that has a coverage of historical events from 1995 to 2015, with 14,757,915 records.

2.1.3 EventRegistry. The EventRegistry [21] project takes a completely different approach than the coded event databases such as GDELT and ICEWS, and performs event extraction based on a clustering of news articles and event mentions. EventRegistry records are formatted in JSON, and contain a multilingual title and summary text, the number of articles reporting the event, the event date (when the event has happened or will happen and not the report date as in coded event databases), and a set of concepts along with the concept type (e.g. location, person, or "topic") and its Wikipedia URL. At the time of this writing, we have ingested 2,889,497 event records extracted from 33,549,078 news articles from the past two years, with 98,435,900 concept annotations, 42,006,079 similarity links, 772,553 location annotations.

2.1.4 Knowledge Sources. In addition to event data, our system also ingests publicly available RDF knowledge bases to use as a source of reference knowledge. Our current knowledge sources include Wikidata [34], DBpedia [13], YAGO [28], and Freebase [14]. At the time of this writing, we have ingested over 6.3 billion RDF triples, containing over 488 million entities (unique URIs) and over 83 million English label statements. We describe our common ingestion and indexing pipeline next.

2.2 Ingestion

As shown in Figure 3, we have a common ingestion pipeline for both the event databases and knowledge sources that is capable of crawling remote sources, parsing structured relational, semi-strcutured (JSON), and RDF (NTriples) data, cleaning invalid records or statements and applying basic filters (e.g., removing non-English


Cognitive Assistant for the Anal...
kg1.pok.ibm.com:9000/questionBased?q=What+would+be+the+consequen
80%
Search
Cognitive Assistant for the Analyst
Deep Analysis
GKG-Based
Go!
Caracas
Protest
Venezuela

Question Analysis

What would be the consequences of violent Protest in Caracas, Venezuela?

Global Context

DeepSim Analysis

Powered by GDELT

DeepSim Context: Location: Venezuela Location: Caracas Theme: PROTEST

Key People	Organizations	Themes
Jesus Torrealba	Venezuela Supreme Court	Tax Ethnicity Venezuelan
Henrique Capriles	United Socialist Party Of Venezuela	Tax Ethnicity Venezuelans
Nicolas Maduro	National Electoral Council	Tax Political Party Unity Alliance
Eyanir Chinea	Venezuelan National Assembly	Tax Econ Freetradeagreements Mercosur
David Smilde	Central University Of Venezuela	Tax Ethnicity Colombians
Julio Borges	Maduro United Socialist Party Of Venezuela	Tax Political Party Democratic Action Party
Freddy Guevara	Venezuela National Electoral Council	Tax Worldlanguages Bogota
Delcy Rodriguez	Venezuelan Supreme Court	Tax Ethnicity Colombian
Nicholas Maduro	National Assembly Henry Ramos Allup	Tax Terror Group Revolutionary Armed Forces
Lilian Tintori	Venezuelan National Guard	Tax Weapons Tear Gassing
Vladimir Padrino	Venezuela National Assembly	Tax Worldlanguages Amazonas
Corina Pons	Venezuela Congress	Tax Ethnicity Uruguayan
Tibisay Lucena	Venezuelan Embassy	Tax Terror Group National Liberation Army
Maria Corina Machado	Penal Forum	Tax Worldlanguages Temer

(a) DeepSim Results

Index-Based Analysis & Events

Powered by EventRegistry

EventRegistry Context: Caracas Protest Venezuela

Topics		Key Players		Locations	
100	Protest	69	Nicolás Maduro	100	Caracas
87	Government	44	Henrique Capriles Radonski	100	Venezuela
81	President of the United States	40	Hugo Chávez	82	United States
71	Violence	23	Barack Obama	45	Colombia
70	Death	19	John Kerry	44	Washington, D.C.
65	Government of the United Kingdom	18	Leopoldo Löpez	26	Ukraine
64	Demonstration (people)	17	Daniel Ortega	25	Panama
57	Twitter	13	Romeo Miller	25	Cuba
55	Opposition (parliamentary)	13	Diosdado Cabello	21	Organization of American States
54	Leadership	10	Fidel Castro	18	Brazil
54	The Capital	9	Jesus	18	Valencia

217 Events Found

- Venezuela high court releases prominent government opponent
- Tensions up in Venezuela as police arrest opposition leader
- Soldados venezuelanos e estudantes se enfrentam em protesto contra Maduro (Portuguese)
- Venezuela issues new protest rules
- Venezuelan Shortages, Long Lines Spark Violence, Arrests
- Maduro slams U.S. gov't over sanctions against Venezuelan officials
- Dutzende Häftlinge in Venezuela trinken Giftmischung
- Venezuela confirms 13 deaths from drug overdoses in prison disturbance
- Venezuela's government recalls ambassador to Spain
- Venezuela launches disarmament plan
- Jailed Venezuelan police chief freed
- Video Surfaces that Shows Venezuelan Opposition Activists Planning Violence
- Delay in Leopoldo Lopez Trial Causes Unrest in Venezuela and Abroad

(b) Index-Based Results

Figure 2: Question Analysis UI - Concept Discovery Results

labels), and finally storing the data. Our platform is implemented on top of Apache Hadoop and Spark, enabling efficient data processing on a cluster on public or private cloud.

2.3 Curation

We adopt a pay-as-you-go integration approach [18, 24] and perform only a minimal curation by a lightweight mapping of known entities and linking them using a common URI when possible. As for the knowledge sources, our integration point is using the existing Wikipedia URLs given that all our sources are based originally on Wikipedia. We then index all the facts (RDF triples) in a key-value store (powered by Riak [9]) in addition to a document store (powered by SolrCloud [3]) that makes it possible to perform highly efficient fact-based or label-based lookups[1]. We also create an auxiliary unified index of common entities using our mapping strategy that results in a collection of 16,108,676 entities with Wikipedia URLs, each linked with one or more of their Wikidata, DBpedia, YAGO, and Freebase URIs. All the event databases are indexed in

[1] We also store the triples in an RDF store (powered by BlazeGraph [5]) although we do not rely on an RDF store for the work described in this paper.

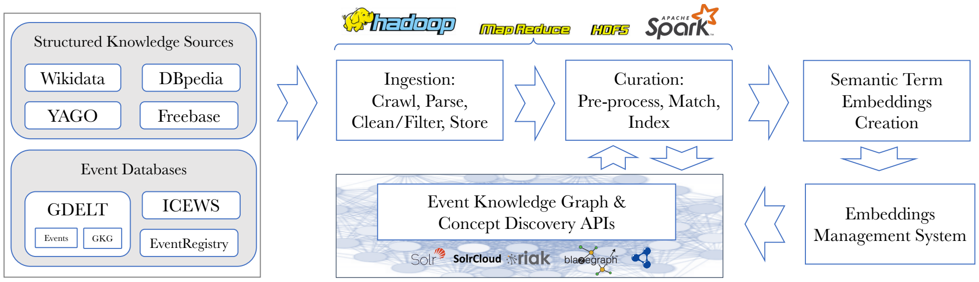

Structured Knowledge Sources
Wikidata
DBpedia
YAGO
Freebase
Event Databases
GDELT
Events
GKG
ICEWS
EventRegistry
hadoop
MapReduce
HDFS
APACHE Spark
Ingestion: Crawl, Parse, Clean/Filter, Store
Curation: Pre-process, Match, Index
Semantic Term Embeddings Creation
Event Knowledge Graph & Concept Discovery APIs
Solr
SolrCloud
riak
blazegraph
Embeddings Management System

Figure 3: System Architecture

a similar way in our key-value and document stores, with labels matched and linked with a Wikipedia URL when possible.

2.4 Semantic Embeddings Engine

Inspired by the idea of word embeddings in NLP [26], recent work has proposed the use of shallow neural networks to map values in structured tables to vectors (referred to as embeddings) [15]. This enables powerful semantic similarity queries even without a prior knowledge of the database schema and contents. We adopt a similar strategy and transform every value in the input event databases into an embeddings vector using a variation of the *continuous skip-gram model* of the original *word2vec* [11, 25, 26]. The first step in this process is a *virtual document* creation process, turning each row in the input database into a context in a corpus of text. This step is performed efficiently in MapReduce. We then feed the text corpus into a word2vec model construction modified to take into account the different characteristics of structured data:

- The order of columns in structured databases is of little importance. While distance between two words in a text document makes them farther in terms of context, the first column in a database table is as relevant to the second column as to the last column.
- In text documents, typically a random-sized window of words is selected. The length of each database record is fixed and so there is no need for a random window size.
- Most importantly, while all words in a text corpus are treated in the same way and do not have specific roles, values in different columns in structured sources describe different (event) features and may need to be grouped and queried differently. There is often a need to search over (or query using) the terms from specific attributes (columns).

Once attribute values are mapped into low-dimensional vectors, aggregate vectors can represent individual records (articles or events), and similarity queries over the vectors can be used for concept discovery and analysis as described in Section 3. These vectors represent the semantic context of every single value seen in the input data, enabling a powerful and extremely efficient method of performing similarity analysis over large amounts of data. As an example, the corpus size (number of words in the "virtual documents") for GDELT GKG is 23,901,358,498 while the size of the vocabulary (number of unique words) in our embeddings is 2,829,213. Still, a key requirement is efficient similarity queries over the vectors with milliseconds running time to enable real-time analysis queries through our UI (Figures 1 & 2) as some analysis queries require several similarity queries each over millions of vectors. We achieve this using the efficient Annoy library [4] as the core of our embeddings management system.

2.5 Event Knowledge Graph & Concept Discovery APIs

The final outcome of all the components is a set of APIs to perform knowledge graph and concept discovery queries. In particular:

- **Lookup APIs** These APIs provide access to the ingested and curated event data and knowledge. For example, one can perform search over knowledge base entity labels and subsequently retrieve human-readable facts as JSON objects. Using this API the user can retrieve infobox-style information about each of the concepts shown under the "Global Context" box in Figure 2a. These APIs also enable queries across event databases, e.g., retrieve ICEWS, GDELT, and EventRegistry events in a given time range that is annotated with a particular concept.
- **Natural Language & Keyword Query Understanding APIs** These APIs turn the user query into a set of knowledge base concepts and event database terms. In Figure 2a, the concepts shown under the "Global Context" are extracted using the API that outputs knowledge base concepts, whereas the terms shown under "DeepSim Context" are terms found in GDELT GKG data used for the shown concept discovery results.
- **Concept Discovery & Ranking APIs** These APIs take a set of concepts or terms and return as output a ranked list of concepts of different types (e.g., Persons, Organizations, Themes). Details of the concept ranking algorithms are described in the following section.

3 CONCEPT RANKING ALGORITHMS

In this section, we describe three classes of algorithms for concept discovery and ranking. These algorithms *identify* and *rank* a set of most relevant concepts of various types (e.g., persons or organizations) for a given set of concepts. An example use of these rankings is shown in Figure 2 where sorted lists of ranked concepts relevant to the user's analysis question are shown. The end goal is providing the output either directly to an analyst or to other components of a decision analysis system, in order to assist with writing a comprehensive and unbiased report on the input concepts based on historical and ongoing events. In what follows, we describe the algorithms and the intuition behind them and in the following section we describe an evaluation framework to measure the effectiveness of these algorithms in identifying relevant concepts.

3.1 Index-Based Method (co-occur)

We refer to this method as the co-occur method. It relies on an efficient index to measure the level of co-occurrence of concepts in a collection of events and uses this as a measure of relevance. Using the index described in Section 2.3, we can search for (all or recent) event records annotated with a given set of concepts. By counting the concept annotations for every record in the output, a list of most frequently co-occurring concepts of various types is returned along with the percentage of co-occurrence of the annotations among all the retrieved event records. Figure 2b shows an example of ranked "Topic", "Key Player" (Person), and "Location" concepts over EventRegistry event records where the API retrieves a maximum of 500 most recent events (an input parameter of the API). The concepts extracted from the input question are "Caracas", "Protest", and "Venezuela". Obviously, these concepts themselves are on top of the lists as they appear in 100% of the event records containing them. The topic concept "Government" appears in 87% of the events and "Nicolás Maduro" appears in 69% of the events, indicating that these concepts are highly relevant to the input concepts in recent events.

3.2 Deep Similarity Method Using Semantic Embeddings (context)

We refer to this as the context method. It relies on the term embeddings built over an event database as described in Section 2.4. First, a vector is retrieved for each of the terms extracted from the input question (where there exists a vector representation in the embeddings space), and an average vector is constructed by summing the values in each dimension and normalizing the resulting vector. Using the embeddings management system, the most similar vectors of various kinds of terms are retrieved, ranked by their similarity to the average vector. Figure 2a shows an example of concept rankings with the same question. The API used in this example queries embeddings built over GDELT GKG, with vectors of size 200 and cosine similarity as our choice of vector similarity measure. Unlike the index-based results above, these rankings reflect the relevance according to the context of these terms in over 157 million GKG records, and result in less-obvious and harder-to-explain but deeply relevant sets of concepts in the output.

3.3 Combination Methods

The co-occur method has the disadvantage of requiring co-occurrence and high-quality concept annotations, and the context method has the disadvantage of results that are hard to explain and often relevant but less important concepts than the co-occur method. As a result, one can combine the two methods as a way of addressing each individual method's limitation. We implement two such combination methods. In the first one, we retrieve a set of 3*k results of an index-based method, re-rank the output using the embeddings-based similarity of the terms in the output, then select the top k terms. We refer to this method as co-occur_context. In the second variation, context_co-occur, we retrieve 3*k results of the context retrieval and sort the output based on their position in the co-occur results before selecting the top k terms. In the following section in Table 1, we show an example of how these re-rankings improve the results.

4 EXPERIMENTS

To evaluate the performance of the concept ranking algorithms, we define a *key player* as a person or organization directly involved with the topic of a query. The strongest form of involvement is to be a direct participant in the current action, or a key player in bringing about the current situation. A lesser form is to make a public statement about the action, or report on the action, and weakly involved players are those used as a comparison or analogy to the current situation. Using this definition, we sought to identify a 'ground truth' of concepts related to a given query, where a concept is a key player. We focus on the performance of the algorithms given a set of query terms that would be extracted from an analyst's question.

Note that here we focus on the evaluation of accuracy. As mentioned in Section 2, our framework enables real-time or near real-time response time for each of the algorithms and so we do not compare running times.

4.1 Evaluation Data

To our knowledge, there is no existing public data set that identifies key people and organizations for a set of analytical questions. To provide an objective basis for our evaluation, we used reports that summarize a political or social event or situation, making the assumption that these reports are a response to such a question, and will mention the most important key players. We did not use news articles because these are the source of GDELT events, so as not to bias the results. We identified three potential sources of reports:

- Declassified US Government intelligence reports. We were only able to identify one such report that relates to the time period covered by the GDELT GKG event database: 'Assessing Russian Activities and Intentions in Recent US Elections', released in January 2017.
- Wikipedia pages describing a newsworthy event or topic with relevance to social unrest. For example 'Impeachment of Dilma Rousseff' or 'Shortages in Venezuela'.
- Human Rights Watch reports. These are detailed descriptions of specific human rights situations around the world, for example 'Philippine Police Killings in Duterte's "War on Drugs"'. 1,091 such reports are available in HTML.

Using these sources we developed test queries consisting of a small set of query terms and 'ground truth' sets of people and organizations. The query terms can include a country, people, organizations, and themes (drawn from the GDELT GKG themes described earlier). The 'ground truth' items are selected from the people and organizations mentioned in the report, to represent the ideal response of the system to the query. Three query sets were developed: *Manual*, *Curated*, and *Auto*, representing different levels of accuracy and ease of production.

4.1.1 Manual. A small set of 12 hand-built queries derived from 6 Wikipedia pages, 5 Human Rights Watch Reports from 2014 or later; and 1 declassified intelligence report. All queries specified the country most strongly associated with the report, and 1-3 manually selected themes from GDELT GKG. Five queries included a person mentioned in the report title, and one specified two organizations. For example, the query terms for the Wikipedia page 'Impeachment of Dilma Rousseff' consisted of the country 'Brazil', the person 'Dilma Rousseff', and the theme 'IMPEACHMENT'.

To define the ground truth concepts, the sets of people and organizations mentioned in each report were ordered using a combination of their importance to the topic, frequency of appearance, and order of appearance in the report (earlier is better). Only the mentioned people and organizations that were key players in the overall topic were included, based on human judgement. For example, if an article drew a comparison between a current leader and a historical figure, the historical figure would not be considered central to the topic itself. We then removed items that were found in the report but not in our embedding. Subsequently, each query had an average of 10 ground truth people and 7 ground truth organizations.

4.1.2 Curated. A set of 25 queries based on Wikipedia pages describing events from 2014-2016, where the query terms (country, people, organizations and themes) were selected manually, but the ground truth terms were automatically generated from the Wikipedia links within the page, and then curated to remove non-person and non-organization terms. Some people and organizations mentioned in the original report may be missed in this process.

4.1.3 Auto. A larger set of 179 queries derived from the Human Rights Watch reports, with fully automatic generation of both query terms and ground truth. To generate the query terms, the query builder used the document title, subtitle and teaser - a short paragraph of a few sentences describing the report. It used concept extraction software that combines output from ClearNLP [6] and OpenNLP [2] to identify noun phrases referring to named entities, and assigns types to them according to their linguistic context. We relied on these types to identify countries, people and organizations in the text. This method introduces noise, with some people or organizations being missed, and other spurious ones found. We leave selection of appropriate query themes to future work, as the theme extraction generates overly specific queries and degrades the results for this query set. The ground truth people and organizations were generated by using the same concept extraction software, applied to the full text of the report. Many of the interviewees in Human Rights Watch reports choose to speak anonymously or use pseudonyms. Thus, we removed people and organizations not found in our embedding. Finally, we selected the 179 queries that had a country, at least one other query term (person, organization or theme), and had ground truth terms that were not already present in the query. Of these, the majority (102) were queries consisting of a country and the single organization 'Human Rights Watch'. 26 contained a person, and 51 contained an organization other than Human Rights Watch. From these, we further selected only the usable queries with ground truth items that were not in the query (143 for people and 155 for organizations) These queries had, on average, 21 ground truth people and 32 ground truth organizations.

4.2 Example Results

Table 1 shows the set of ground truth people and the output of the algorithms for a query from the manual test set, based on the 2016 Human Rights Watch report "Venezuela's Humanitarian Crisis. Severe Medical and Food Shortages, Inadequate and Repressive Government Response"[2].

Of the 14 most relevant people identified in the report, only 7 were present in the embedding (indicated in column 1 with (*)). Two of the others were not found in the GKG data at all, and the remaining five were mentioned only 1-59 times - not enough to be included in the embedding. One surprising missing name was Ban Ki-moon, Secretary-General of the United Nations. For organizations, 17 were mentioned in the report, and 9 of these were found in the embedding. Again, some of the missing organizations were surprising, for example, the Organization of American States. These gaps in coverage indicate noise and loss in the GKG event extraction process. The GKG data does not often include common acronyms like BBC or FBI, although there are some exceptions. This creates challenges for automated testing since the reports often use an acronym to refer to an organization.

The most relevant people mentioned in the report, 14 in all, are listed in the first column of Table 1, while the remaining columns show the top 14 results for each algorithm, given the query for the country "Venezuela" and the GKG theme "SELF_IDENTIFIED_HUMANITARIAN_CRISIS". The seven items from the ground truth that are potentially findable in the index and in the embedding are indicated with (*) in column one, while the found items are highlighted in bold in the subsequent columns, including alternate spellings of the same person's name.

The ground truth list includes the current and former leaders of Venezuela (Nicolás Maduro and Hugo Chávez), the United Nations High Commissioner for Human Rights (Zeid Ra'ad Al Hussein), and Secretary General (Ban Ki-Moon), four Venezuelan politicians and state governors (Delcy Rodríguez, Luisana Melo, Julio León Heredia, Diosdado Cabello), the Secretary-General of the Association of Latin American States (Luis Almagro), three doctors whose stories illustrate the crisis and government response (Johan Gabriel Pinto Graterol, Carlos Zapa, Flor Sánchez), and two human rights leaders (Rafael Uzcátegui, Feliciano Reyna).

The `co-occur` method finds only the Venezuelan leaders in the top 14. It also returns ten other world leaders, politicians and spokespeople (Barack Obama, Rafael Correa, John Kerry, Bashar Assad, Donald Trump, Vladimir Putin, Juan Manuel Santos, David Granger, John Kirby and Salva Kiir. These people have either made statements about Venezuela's humanitarian crisis, or Venezuela has made comments about their own country's crisis (e.g. Bashar Assad). Although Donald Trump was not yet president of the United States during the period covered by the data, his opinions on foreign policy in Latin America were discussed in the news, and he made statements about the situation in Venezuela. Two journalists who write frequently about Venezuela are also suggested (Joshua Goodman, Gonzalo Solano).

[2] https://www.hrw.org/report/2016/10/24/venezuelas-humanitarian-crisis/severe-medical-and-food-shortages-inadequate-and

Table 1: Example results from each algorithm for the query "Venezuela", "SELF_IDENTIFIED_HUMANITARIAN_CRISIS". (*) indicates those candidates that were potentially findable in the GKG data.

Ground truth	`co-occur`	`context`	`co-occur_context`	`context_co-occur`
Nicolás Maduro (*)	**Nicolás Maduro**	**Delcy Rodriguez**	**Delcy Rodriguez**	**Nicolás Maduro**
Hugo Chávez (*)	Barack Obama	**Nicholás Maduro**	**Nicholás Maduro**	Juan Manuel Santos
Zeid Ra'ad Al Hussein (*)	Rafael Correa	Jesus Torrealba	Hannah Dreier	Hannah Dreier
Ban Ki-moon	**Hugo Chávez**	Vladimir Padrino	**Luis Almagro**	**Luis Almagro**
Delcy Rodríguez (*)	Joshua Goodman	Henrique Capriles	Juan Manuel Santos	**Delcy Rodriguez**
Luisana Melo (*)	John Kerry	**Nicolás Maduro**	Barack Obama	**Nicholás Maduro**
Luis Almagro (*)	Bashar Assad	Jorge Arreaza	Rafael Correa	Jesus Torrealba
Johan Gabriel Pinto Graterol	Donald Trump	Hannah Dreier	**Hugo Chávez**	Vladimir Padrino
Julio León Heredia	Juan Manuel Santos	Girish Gupta	Joshua Goodman	Henrique Capriles
Carlos Zapa	Gonzalo Solano	Eyanir Chinea	John Kerry	Jorge Arreaza
Flor Sánchez	Vladimir Putin	Andrew Cawthorne	Bashar Assad	Girish Gupta
Diosdado Cabello (*)	David Granger	David Smilde	Donald Trump	Eyanir Chinea
Rafael Uzcátegui	John Kirby	Ernesto Villegas	Gonzalo Solano	Andrew Cawthorne
Feliciano Reyna	Salva Kiir	**Luis Almagro**	Vladimir Putin	David Smilde

In marked contrast, the `context` method's results do not include any foreign leaders and politicians. Instead, there are seven Venezuelan politicians (Delcy Rodriguez, Nicholás Maduro, Jesus Torrealba, Vladimir Padrino, Henrique Capriles, Jorge Arreaza, and Ernesto Vollegas), along with four journalists (Hannah Dreier, Girish Gupta, Eyanir Chinea and Andrew Cawthorne) and a human rights advocate and academic (David Smilde), and the secretary-general of the Organization of Latin American States (Luis Almagro). Some of these politicians are very closely associated with the humanitarian crisis in Venezuela, notably Vladimir Padrino, the Venezuelan Minister of Defense, who is responsible for food distribution, even though they were not mentioned by name in the report.

Combining these methods by ranking the first 90 `co-occur` results according to their `context` ranking moved five highly related candidates to the top of the list, including a new ground truth person: Luis Almagro. The remainder of the list is composed of the unused candidates from the original `co-occur` list (Barack Obama, etc). Similarly, the `context_co-occur` method moved four items to the top of the ranking, including the misspelling of Nicolás Maduro as Nicholás Maduro. Both combination methods slightly increased the number of ground truth items found in the top 10 ranked results from 2 or 3 to 4 out of a possible maximum of 7.

4.3 Evaluation Method

To evaluate and compare the methods of identifying key players, we applied each of the four methods (`co-occur`, `context`, `co-occur_context` and `context_co-occur`) to the test query data sets (manual, curated and auto), for both people and organizations. All methods were limited to 30 returned candidates. Some queries had no ground truth people, or no organizations, and these tests were skipped. For each query, we calculated four classic information retrieval evaluation measures: precision (ratio of correct concepts in the output), recall (ratio of ground truth concepts returned in the output), F1 (harmonic mean of precision and recall), and average precision (average precision value at all the ranks where a correct concept is returned). Overall values for each test set are reported as the mean of the values for the individual queries in the set. Following the recommendation by Smucker et al. [30], we performed randomization test and two-tailed paired samples t-tests to test for statistical significance.

4.4 Evaluation Results

4.4.1 Manual. Table 2 shows the results for the manual data set. For person experiments, all measures showed better performance from the combination methods, with the `context` method performing the lowest. The `co-occur_context` method outperformed the `co-occur` method by 19%. However, pairwise comparisons of F1 scores between methods showed only the (`context`,`context_co-occur`) and (`co-occur`,`co-occur_context`) pairs to be statistically significantly different (p< 0.05). For the organization experiments, again the co-occur_context combination method performed best over all four measures, but only the (`context`,`context_co-occur`) pair was found to be statistically significant in terms of comparison by MAP or F1 scores. The lack of statistical significance is due to the high variance of the results for each query, and show in part the need for a larger data set for a proper comparison as our overall results described in Section 4.4.4 also confirm.

Table 2: Accuracy results over the manual data set.

	person				organization			
	`co-occur`	`context`	`co-occur` `context`	`context` `co-occur`	`co-occur`	`context`	`co-occur` `context`	`context` `co-occur`
MAP	0.233	0.199	0.251	0.233	0.179	0.143	0.184	0.189
F1	0.192	0.174	0.228	0.213	0.178	0.107	0.183	0.141
Pr.	0.133	0.121	0.158	0.149	0.117	0.066	0.119	0.089
Re.	0.372	0.328	0.437	0.388	0.436	0.304	0.459	0.374

4.4.2 Curated. Table 3 shows the results for the curated data set. For the person experiments, the overall pattern was very similar to the manual data set, with the `co-occur_context` method showing the best performance across all measures, including an 18% improvement for F1 over the `co-occur` method. For F1, the differences between the (`context`,`co-occur`), (`context`,`context_co-occur`) and (`co-occur`,`co-occur_context`) pairs were statistically significant. For the organization experiments the `co-occur` and `co-occur_context` methods performed the best, and their F1 scores were not significantly different, while all other pairwise comparisons were, except for the two lowest performing methods: `context` and `context_co-occur`.

4.4.3 Auto. Table 4 shows the results for the auto data set. For these results, organizations followed a similar pattern to the two other datasets, and all pairwise comparisons were statistically significant for all metrics, with the only exception for MAP, where the two combination methods were not distinguishable. For person

Table 3: Accuracy results over the curated data set.

	person				organization			
	co-occur	context	co-occur context	context co-occur	co-occur	context	co-occur context	context co-occur
MAP	0.132	0.070	0.135	0.123	0.130	0.039	0.075	0.051
F1	0.140	0.104	0.165	0.143	0.142	0.058	0.142	0.058
Pr.	0.119	0.090	0.139	0.124	0.107	0.042	0.108	0.045
Re.	0.251	0.160	0.300	0.225	0.290	0.116	0.289	0.122

experiments, the results were lower, less than 0.1 for all metrics and methods, so that while the combination methods produced around 10% higher average scores, the differences were not statistically significant, with the exception of the (context,context_co-occur) and (co-occur,co-occur_context) pairs for F1 or MAP.

Table 4: Accuracy results over the auto data set.

	person				organization			
	co-occur	context	co-occur context	context co-occur	co-occur	context	co-occur context	context co-occur
MAP	0.041	0.046	0.050	0.051	0.132	0.073	0.117	0.116
F1	0.058	0.060	0.066	0.066	0.165	0.084	0.157	0.108
Pr.	0.051	0.056	0.059	0.061	0.173	0.088	0.163	0.112
Re.	0.090	0.086	0.099	0.094	0.224	0.114	0.217	0.150

4.4.4 Comparing the data sets. We also explored whether the different data sets provided similar results. Table 5 summarizes the mean average precision (MAP), F1, precision and recall values for each data set, for people and organizations, averaging over all methods, while Figure 4 shows F1 values for people (left) and organizations (right) as boxplots, for each data set and method. Each box indicates the interquartile range of the data, the center line indicates the median value, the whiskers above and below give the 95% confidence intervals, and circles indicate outliers. Significant differences are indicated above with red brackets. For people, the less curated sets of queries produced lower results, but the pattern of results is very similar across all three datasets. Recall that the auto queries did not contain any themes, and so they often did not capture the topic of a report well, giving the system a low chance of success. Results were twice as good for organizations as for people in the auto data set, probably reflecting the large number of queries that included an organization. Again, the pattern of results remained similar across the data sets.

Table 5: Comparison of mean metrics for people and organizations in the manual, curated and auto test query sets.

	person			organization		
	manual	curated	auto	manual	curated	auto
MAP	0.23	0.12	0.05	0.17	0.07	0.11
F1	0.20	0.14	0.06	0.15	0.10	0.13
Pr.	0.14	0.12	0.06	0.10	0.08	0.13
Re.	0.38	0.23	0.09	0.39	0.20	0.18

4.5 Discussion

Overall, both the combination algorithms performed better than the individual co-occur and context algorithms. This suggests that combining methods did go some way towards addressing the individual weaknesses of the two approaches, with an effect size of up to 19% improvement over the next best individual method.

Not surprisingly, the algorithms produced the best results on the manual test set, followed by the curated set, and the lowest values for the automatically generated set, which has less well constructed queries that do not capture the topic of the report as well. Importantly, the similarities between data set results when comparing the concept discovery algorithms increases confidence in the evaluation, and more generally in the use of automated methods as a valid and scalable way to approach the evaluation of concept discovery algorithms, despite the noise and loss of accuracy compared to hand-curated data.

Our approach to evaluation has some limitations. Our source reports do not mention all of the people and organizations relevant to the topic by name. We do not translate mentions like "The Minister for the Interior" into a named person, and nor do we attempt to resolve references to groups like "Brazilian steel companies." Our methods also draw from articles published after the publication of the report, when new concepts may be introduced. All people or organizations found by our methods but not named in the ground truth are treated as wrong answers, but some of these may be highly relevant to the topic. In the example shown in Table 1, the majority of the persons returned by the context method are in fact highly relevant despite the fact that our input report did not contain their names. This shows that a potential use case for our system is complementing analysts in finding concepts that are not already covered in their report. Also note that there is often a major difference between the number of candidates proposed (30) and the number of ground truth items provided (generally less than 30). Thus, our reported precision (and therefore F1) scores may seriously underestimate the quality of the responses.

We selected Human Rights Watch reports because they tackle the kinds of analysis questions we are interested in. To address the problem of relevant concepts not being mentioned, we will explore whether there are better sources of reports and other sources similar to Wikipedia, for the auto method, given that Wikipedia articles are intended to provide an overview of the topic. We will also examine how to automatically select the most appropriate small set of themes for queries. As a means to compare concept extraction algorithms, our approach of combining high quality hand-built test queries with a larger, less precise query set provides a balance between quality and quantity and, we argue, is effective only as a means of comparing different concept extraction algorithms. To get a more accurate measure of overall precision, human judgement could be used to construct more complete test queries, or to evaluate whether each proposed candidate is relevant to a question. These solutions would be labor-intensive, but could be suitable for crowdsourcing approaches.

5 CONCLUSION

In this paper, we presented a framework for discovering concepts related to an analysis question using event databases. We discussed the system's architecture and implementation details of its components. We presented three classes of concept ranking algorithms and evaluated the quality of the rankings using a corpus of reports written by humans, and based on both manual and automatically generated ground truth questions and concepts. Our evaluation shows promising results for our semantic embeddings based concept ranking particularly when combined with an index-based co-occurrence based ranking. We also presented challenges in our evaluation and a few avenues for future work.

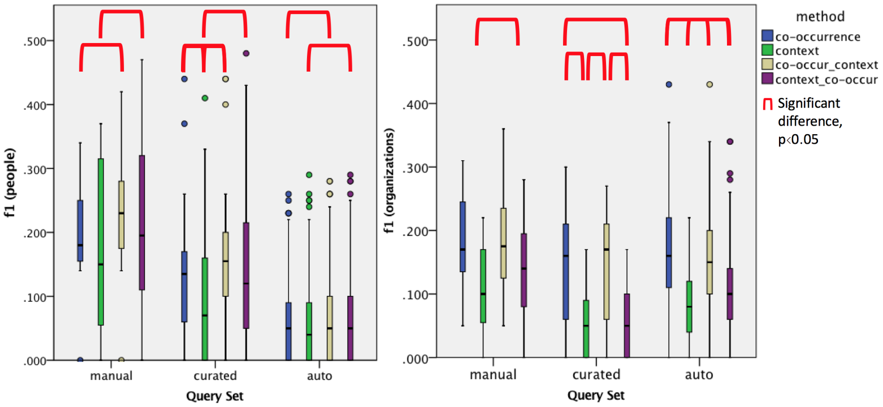


Figure 4: Results for people and organizations across the three data sets and four methods

REFERENCES

[1] AlchemyData News | IBM Watson Developer Cloud. https://www.ibm.com/watson/developercloud/alchemy-data-news.html. [Online; accessed May 8, 2017].

[2] Apache OpenNLP (Version 1.5.3). https://opennlp.apache.org/. [Online; accessed May 18, 2017].

[3] Apache Solr: SolrCloud. https://wiki.apache.org/solr/SolrCloud. [Online; accessed May 8, 2017].

[4] Approximate Nearest Neighbors in C++/Python optimized for memory usage and loading/saving to disk. https://github.com/spotify/annoy. [Online; accessed May 8, 2017].

[5] Blazegraph. https://www.blazegraph.com. [Online; accessed May 8, 2017].

[6] ClearNLP (Version 3.2.0). https://github.com/clir/clearnlp. [Online; accessed May 18, 2017].

[7] Google News. https://news.google.com/. [Online; accessed May 8, 2017].

[8] Introducing the Global Content Analysis Measures (GCAM). http://blog.gdeltproject.org/introducing-the-global-content-analysis-measures-gcam/. [Online; accessed May 8, 2017].

[9] Riak: a distributed, decentralized data storage system. https://github.com/basho/riak. [Online; accessed May 8, 2017].

[10] TAC KBP 2016 Event Track. https://tac.nist.gov/2016/KBP/Event/index.html. [Online; accessed May 8, 2017].

[11] word2vec: Tool for computing continuous distributed representations of words. https://code.google.com/archive/p/word2vec/. [Online; accessed May 8, 2017].

[12] World-Wide Integrated Crisis Early Warning System. http://www.lockheedmartin.com/us/products/W-ICEWS.html. [Online; accessed May 8, 2017].

[13] C. Bizer, J. Lehmann, G. Kobilarov, S. Auer, C. Becker, R. Cyganiak, and S. Hellmann. 2009. DBpedia - A Crystallization Point for the Web of Data. *J. Web Sem.* 7, 3, 154–165.

[14] K. D. Bollacker, C. Evans, P. Paritosh, T. Sturge, and J. Taylor. 2008. Freebase: a collaboratively created graph database for structuring human knowledge. In *ACM SIGMOD Int'l Conf. on Mgmt. of Data*. 1247–1250.

[15] Rajesh Bordawekar and Oded Shmueli. 2016. Enabling Cognitive Intelligence Queries in Relational Databases using Low-dimensional Word Embeddings. *CoRR* abs/1603.07185. http://arxiv.org/abs/1603.07185

[16] Elizabeth Boschee, Jennifer Lautenschlager, Sean O'Brien, Steve Shellman, James Starz, and Michael Ward. 2017. ICEWS Coded Event Data. https://doi.org/10.7910/DVN/28075

[17] G. Doddington, A. Mitchell, M. Przybocki, L. Ramshaw, S. Strassel, and R. Weischedel. 2004. The Automatic Content Extraction (ACE) Program Tasks, Data, and Evaluation. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC-2004)*. European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2004/pdf/5.pdf ACL Anthology Identifier: L04-1011.

[18] Michael J. Franklin, Alon Y. Halevy, and David Maier. 2005. From databases to dataspaces: a new abstraction for information management. *SIGMOD Record* 34, 4, 27–33. https://doi.org/10.1145/1107499.1107502

[19] Frederik Hogenboom, Flavius Frasincar, Uzay Kaymak, Franciska de Jong, and Emiel Caron. 2016. A Survey of Event Extraction Methods from Text for Decision Support Systems. *Decis. Support Syst.* 85, C, 12–22. https://doi.org/10.1016/j.dss.2016.02.006

[20] Gizem Korkmaz, Jose Cadena, Chris J. Kuhlman, Achla Marathe, Anil Vullikanti, and Naren Ramakrishnan. 2015. Combining Heterogeneous Data Sources for Civil Unrest Forecasting. In *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015 (ASONAM '15)*. 258–265. http://doi.acm.org/10.1145/2808797.2808847

[21] Gregor Leban, Blaz Fortuna, Janez Brank, and Marko Grobelnik. 2014. Event Registry: Learning About World Events from News. In *Proceedings of the 23rd International Conference on World Wide Web (WWW '14 Companion)*. 107–110. https://doi.org/10.1145/2567948.2577024

[22] Kalev Leetaru and Philip A Schrodt. 2013. GDELT: Global data on events, location, and tone, 1979–2012. In *ISA Annual Convention*.

[23] Dekang Lin and Patrick Pantel. 2002. Concept Discovery from Text. In *Proceedings of the 19th International Conference on Computational Linguistics - Volume 1 (COLING '02)*. 1–7. https://doi.org/10.3115/1072228.1072372

[24] Jayant Madhavan, Shawn R Jeffery, Shirley Cohen, Xin Dong, David Ko, Cong Yu, and Alon Halevy. 2007. Web-scale data integration: You can only afford to pay as you go. In *Proceedings of CIDR*. 342–350.

[25] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

[26] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed Representations of Words and Phrases and Their Compositionality. In *Advances in neural information processing systems*. 3111–3119.

[27] S. Muthiah, P. Butler, R. P. Khandpur, P. Saraf, N. Self, A. Rozovskaya, L. Zhao, J. Cadena, C. Lu, A. Vullikanti, A. Marathe, K. Summers, G. Katz, A. Doyle, J. Arredondo, D. K. Gupta, D. Mares, and N. Ramakrishnan. 2016. EMBERS at 4 Years: Experiences Operating an Open Source Indicators Forecasting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 205–214. https://doi.org/10.1145/2939672.2939709

[28] Thomas Rebele, Fabian M. Suchanek, Johannes Hoffart, Joanna Biega, Erdal Kuzey, and Gerhard Weikum. 2016. YAGO: A Multilingual Knowledge Base from Wikipedia, Wordnet, and Geonames. In *The Semantic Web - ISWC 2016 - 15th International Semantic Web Conference*. 177–185.

[29] Philip A Schrodt, Omür Yilmaz, Deborah J Gerner, and Dennis Hermreck. 2008. The CAMEO (conflict and mediation event observations) actor coding framework. In *2008 Annual Meeting of the International Studies Association*.

[30] Mark D. Smucker, James Allan, and Ben Carterette. 2007. A Comparison of Statistical Significance Tests for Information Retrieval Evaluation. In *Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management (CIKM '07)*. 623–632. http://doi.acm.org/10.1145/1321440.1321528

[31] Shirin Sohrabi, Anton V. Riabov, and Octavian Udrea. 2017. State Projection via AI Planning. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*. 4611–4617. http://aaai.org/ocs/index.php/AAAI/AAAI17/paper/view/14277

[32] Shirin Sohrabi, Octavian Udrea, Anton V. Riabov, and Oktie Hassanzadeh. 2016. Interactive Planning-Based Hypothesis Generation with LTS++. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, IJCAI 2016*. 4268–4269. http://www.ijcai.org/Abstract/16/654

[33] Ah-hwee Tan. 1999. Text Mining: The state of the art and the challenges. In *In Proceedings of the PAKDD 1999 Workshop on Knowledge Disocovery from Advanced Databases*. 65–70. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.132.6973

[34] D. Vrandecic and M. Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Commun. ACM* 57, 10, 78–85.

[35] Michael D Ward, Andreas Beger, Josh Cutler, Matt Dickenson, Cassy Dorff, and Ben Radford. 2013. Comparing GDELT and ICEWS event data. *Analysis* 21, 267–297.